Title: Exome chip meta-analysis fine maps causal variants and elucidates the genetic architecture of rare coding variants in smoking and alcohol use.

Running Title: Exome Meta-Analysis of Smoking and Alcohol

Keywords: Tobacco, Nicotine, Alcohol, GWAS, Heritability, Behavioral Genetics

Number of words in abstract: 249
Number of words in main text: 3676
Number of Figures: 0
Number of Tables: 4
Number of Supplemental Materials: One Supplementary Note with eight supplementary tables and four supplementary figures.

Authors, in order with affiliation:

David M. Brazel*	Institute for Behavioral Genetics, University of Colorado Boulder Department of Molecular, Cellular, and Developmental Biology, University of Colorado Boulder
Yu Jiang*	Department of Public Health Sciences, Penn State College of Medicine, Hershey, PA
Jordan M. Hughey*	Department of Public Health Sciences, Penn State College of Medicine, Hershey, PA
Valérie Turcot	Montreal Heart Institute, Montreal, Quebec, H1T 1C8, Canada Department of Medicine, Faculty of Medicine, Université de Montréal, Montreal, Quebec, H3T 1J4, Canada
Xiaowei Zhan	Department of Clinical Science, Center for Genetics of Host Defense, University of Texas Southwestern
Jian Gong	Public Health Sciences Division, Fred Hutchinson Cancer Research Center
Chiara Batini	Department of Health Sciences, University of Leicester
J. Dylan Weissenkampen	Department of Public Health Sciences, Penn State College of Medicine, Hershey, PA
MengZhen Liu	Department of Psychology, University of Minnesota

CHD Exome+ Consortium[††]
Consortium for Genetics of Smoking Behaviour[††]

Daniel R. Barnes	Department of Public Health and Primary Care, University of Cambridge, Cambridge, UK
Sarah Bertelsen	Department of Neuroscience, Icahn School of Medicine at Mount Sinai
Yi-Ling Chou	Washington University
A. Mesut Erzurumluoglu	Department of Health Sciences, University of Leicester

Jessica D. Faul	Survey Research Center, Institute for Social Research, University of Michigan
Jeff Haessler	Public Health Sciences Division, Fred Hutchinson Cancer Research Center
Anke R. Hammerschlag	Department of Complex Trait Genetics, Center for Neurogenomics and Cognitive Research, Amsterdam Neuroscience, VU University Amsterdam
Chris Hsu	University of Southern California
Manav Kapoor	Department of Neuroscience, Icahn School of Medicine at Mount Sinai
Dongbing Lai	Department of Medical and Molecular Genetics, Indiana University School of Medicine
Nhung Le	Department of Medical Microbiology, Immunology and Cell Biology, Southern Illinois University School of Medicine
Christiaan A de Leeuw	Department of Complex Trait Genetics, Center for Neurogenomics and Cognitive Research, Amsterdam Neuroscience, VU University Amsterdam
Anu Loukola	Institute for Molecular Medicine Finland (FIMM), University of Helsinki, Helsinki, Finland; Department of Public Health, University of Helsinki, Helsinki, Finland
Massimo Mangino	Twin Research & Genetic Epidemiology Unit, Kings College, London
Carl A. Melbourne	Department of Health Sciences, University of Leicester
Giorgio Pistis	Istituto di Ricerca Genetica e Biomedica, Consiglio Nazionale delle Ricerche (CNR), Monserrato, Cagliari, Italy.
Beenish Qaiser	Institute for Molecular Medicine Finland (FIMM), University of Helsinki, Helsinki, Finland; Department of Public Health, University of Helsinki, Helsinki, Finland
Rebecca Rohde	University of North Carolina, Chapel Hill
Yaming Shao	University of North Carolina, Chapel Hill
Heather Stringham	Department of Biostatistics, University of Michigan
Leah Wetherill	Department of Medical and Molecular Genetics, Indiana University School of Medicine
Wei Zhao	Department of Epidemiology, School of Public Health, University of Michigan
Arpana Agrawal	Department of Psychiatry, Washington University School of Medicine
Laura Bierut	Department of Psychiatry, Washington University School of Medicine
Chu Chen	Public Health Sciences Division, Fred Hutchinson Cancer Research Center Department of Epidemiology and Department of Otolaryngology; Head and Neck Surgery, University of Washington, Seattle, WA

Charles B. Eaton	Department of Family Medicine, Brown University, Providence, RI
Alison Goate	Department of Neuroscience, Icahn School of Medicine at Mount Sinai
Christopher Haiman	Department of Preventative Medicine, Keck School of Medicine, University of Southern California
Andrew Heath	Department of Psychiatry, Washington University
William G. Iacono	Department of Psychology, University of Minnesota
Nicholas G. Martin	Queensland Institute for Medical Research
Tinca J. Polderman	Department of Complex Trait Genetics, Center for Neurogenomics and Cognitive Research, Amsterdam Neuroscience, VU University Amsterdam
Alex Reiner	Public Health Sciences Division, Fred Hutchinson Cancer Research Center Department of Epidemiology, University of Washington, Seattle, WA
John Rice	Departments of Psychiatry and Mathematics, Washington University St. Louis
David Schlessinger	National Institute on Aging, National Institutes of Health
H Steven Scholte	Department of Psychology, University of Amsterdam & Amsterdam Brain and Cognition, University of Amsterdam
Jennifer A. Smith	Department of Epidemiology, School of Public Health, University of Michigan
Jean-Claude Tardif	Montreal Heart Institute, Montreal, Quebec, H1T 1C8, Canada Department of Medicine, Faculty of Medicine, Université de Montréal, Montreal, Quebec, H3T 1J4, Canada
Hilary A. Tindle	Department of Medicine, Vanderbilt University, Nashville, TN
Andreis R van der Leij	Department of Psychology, University of Amsterdam & Amsterdam Brain and Cognition, University of Amsterdam
Michael Boehnke	Department of Biostatistics, School of Public Health, University of MIchigan
Jenny Chang-Claude	Division of Cancer Epidemiology, German Cancer Research Center
Francesco Cucca	Istituto di Ricerca Genetica e Biomedica, Consiglio Nazionale delle Ricerche (CNR), Monserrato, Cagliari, Italy.
Sean P. David	Department of Medicine, Stanford University, Stanford, CA
Tatiana Foroud	Department of Medical and Molecular Genetics, Indiana University School of Medicine
Joanna M. Howson	Department of Public Health and Primary Care, University of Cambridge
Sharon L.R. Kardia	Department of Epidemiology, School of Public Health, University of Michigan
Charles Kooperberg	Public Health Sciences Division, Fred Hutchinson Cancer Research Center
Markku Laakso	University of Eastern Finland, Finland
Guillaume Lettre	Montreal Heart Institute, Montreal, Quebec, H1T 1C8, Canada

	Department of Medicine, Faculty of Medicine, Université de Montréal, Montreal, Quebec, H3T 1J4, Canada
Pamela Madden	Department of Psychiatry, Washington University
Matt McGue	Department of Psychology, University of Minnesota
Kari North	Department of Epidemiology, University of North Carolina, Chapel Hill
Danielle Posthuma	Department of Complex Trait Genetics, Center for Neurogenomics and Cognitive Research, Amsterdam Neuroscience, VU University Amsterdam Department of Clinical Genetics, VU University Medical Centre Amsterdam, Amsterdam Neuroscience
Timothy Spector	Department of Genetic Epidemiology, Kings College, London
Daniel Stram	Department of Preventative Medicine, Keck School of Medicine, University of Southern California
Martin D. Tobin	Department of Health Sciences, University of Leicester
David R. Weir	Survey Research Center, Institute for Social Research, University of Michigan
Jaakko Kaprio	Institute for Molecular Medicine Finland (FIMM), University of Helsinki, Helsinki, Finland; Department of Public Health, University of Helsinki, Helsinki, Finland
Gonçalo R. Abecasis	Department of Biostatistics, School of Public Health, University of Michigan
Dajiang J. Liu†	Institute of Personalized Medicine, Penn State College of Medicine
Scott Vrieze†	Department of Psychology, University of Minnesota

*These authors contributed equally to the work.

†Address correspondence to Scott Vrieze (vrieze@umn.edu), University of Minnesota, 75 East River Road, Minneapolis, MN 55455; or Dajiang J. Liu (dajiang.liu@psu.edu), Penn State College of Medicine, HCAR 2020, Hershey, PA.

††See supplement for a list of authors associated with the replication consortia.

Abstract:

Background: Smoking and alcohol use have been associated with common genetic variants in multiple loci. Rare variants within these loci hold promise in the identification of biological mechanisms in substance use. Exome arrays and genotype imputation can now efficiently genotype rare nonsynonymous and loss of function variants. Such variants are expected to have deleterious functional consequences, and contribute to disease risk.

Methods: We analyzed ~250,000 rare variants from 16 independent studies genotyped with exome arrays and augmented this dataset with imputed data from the UK Biobank. Associations were tested for five phenotypes: cigarettes per day, pack years, smoking initiation, age of smoking initiation, and alcoholic drinks per week. We conducted stratified heritability analyses, single-

variant tests, and gene-based burden tests of nonsynonymous/loss of function coding variants. We performed a novel fine mapping analysis to winnow the number of putative causal variants within associated loci.

Results: Meta-analytic sample sizes ranged from 152,348-433,216, depending on the phenotype. Rare coding variation explained 1.1-2.2% of phenotypic variance, reflecting 11%-18% of the total SNP heritability of these phenotypes. We identified 171 genome-wide associated loci across all phenotypes. Fine mapping identified putative causal variants with double base-pair resolution at 24 of these loci, and between 3 and 10 variants for 65 loci. 20 loci contained rare coding variants in the 95% credible intervals.

Conclusions: Rare coding variation significantly contributes to the heritability of smoking and alcohol use. Fine mapping GWAS loci identifies specific variants contributing to the biological etiology of substance use behavior.

Introduction

Tobacco and alcohol use together account for more morbidity and mortality in Western society than any other single risk factor or health condition(1). These preventable and modifiable behaviors are heritable(2), but previous human and model organism research, including genome-wide association studies of common variants, have resulted in few associated genetic variants, which most prominently feature genes involved in alcohol/nicotine metabolism and nicotinic receptors(3-7).

Advances in sequencing, genotyping, and genotype imputation now allow cost effective investigation of rare and low frequency variants. Compared to common variants (minor allele frequency [MAF] > 1%) most commonly used in genome-wide association studies (GWAS), rare variants have greater potential to elucidate biological mechanisms of complex traits, including substance use and addiction(8, 9). In particular, nonsynonymous and loss of function (LoF) coding variants, which result in the loss of normal function of a protein, may have greater phenotypic impact and more direct mechanistic interpretation than other variants that do not have obvious biological consequences(10, 11).

No large-scale genome- or exome-wide study of rare variation has been conducted to date. The vast majority of existing addiction-related rare variant studies have used targeted sequencing of putative addiction-associated loci to discover and test for association in relatively small samples. Existing research has led to intriguing leads, including rare variant associations in loci that span nicotinic receptor gene clusters(12-21) and alcohol metabolism genes(22-24) for nicotine and alcohol dependence, respectively. This strategy has also produced rare variant associations in novel loci. In one case, gene-level association tests were used to find an association with rare variants in *SERINC2*(24). In another case, a burden test across *PTP4A1*, *PHF3*, and *EYS* showed association with alcohol dependence(25). Unfortunately, these genes are not obviously involved in etiological processes related to addiction, and replications have not been reported to date.

Previous studies have also attempted to leverage information about predicted functional consequences of rare mutations to improve association analyses. One study of nicotine dependence found significant rare single-variant associations in *CHRNB4*, but only when variants were weighted by their predicted effect on the cellular response to nicotine and acetylcholine(26). Such positive findings could benefit from replication, which has not always been straightforward. For example, all rare variant associations in addiction are, to our knowledge, candidate gene analyses with type I error thresholds based only on the number of tests within that region. Historically, such analyses have produced overly optimistic estimates of the number of associated loci(27). Genome-wide analyses with more conservative type I error thresholds have reported null rare variant findings across an array of phenotypes relevant to addiction(28-30). Precisely because genome-wide analyses are conducted on many variants across the genome, they are in principle able to discover novel rare variant associations within new or known loci. One way to improve power in genome-wide analyses is through genetic association meta-analysis, which entails the aggregation of results across many studies to achieve large sample sizes.

Here, we attempted to expand on these previous discoveries by conducting the largest meta-analytic investigation of exonic rare variants to date. We conducted an exome-wide association meta-analysis of nicotine and alcohol use across 16 studies genotyped on the exome array, which genotypes low-frequency nonsynonymous and putative loss of function exonic variants. We combined these data with the UK Biobank, which includes approximately 400,000 individuals of European ancestry with genotype imputation to the Haplotype Reference Consortium(31) imputation reference panel and relevant smoking/drinking phenotypes. Sample sizes for well-imputed variants were thus enlarged and the availability of noncoding variants from UK Biobank enabled comprehensive analysis of genetic architecture(32) and fine mapping(33).

We conducted single variant and gene-based tests of association with five smoking and drinking phenotypes. We applied a novel fine mapping analysis to prioritize causal variants using statistical and functional information. We also evaluated the contribution of rare exonic variants

to the heritability of these phenotypes. Family studies, as well as studies of the aggregate effects of common variants, have found both alcohol use and tobacco use to be heritable behaviors(30, 34-38). Research on the aggregate contribution of rare variants, however, has been scarce, with previous work on related phenotypes in smaller samples failing to detect aggregate effects for smoking and alcohol consumption(28). We used meta-analytic summary statistics to quantify the contribution to heritability of variants in various functional categories and frequency bins.

Methods and Materials

Seventeen studies contributed summary statistics for meta-analysis. These studies, their sample sizes, and available phenotypes are listed in the online supplement (**Tables S1 and S2**). We augmented our sixteen exome chip cohorts with the UK Biobank, where imputation to the Haplotype Reference Consortium panel was used in lieu of an exome chip array. All individuals were of European ancestry, as determined by genetic principal components.

Phenotypes

Phenotypes were selected to represent multiple stages of smoking. These included initiation, heaviness of use among smokers, and a measure of total lifetime exposure to tobacco. For alcohol use only a measure of amount of alcohol use was systematically available across studies. The selected phenotypes are relevant to prior GWAS of smoking and alcohol use; are commonly available in psychological, medical, and epidemiological data sets; and are known to be correlated with measures of substance dependence(4, 39-41).

1. Cigarettes per day (CigDay). The average number of cigarettes smoked in a day among current and former smokers. Studies with binned responses used their existing bins. Studies that recorded an integer value binned responses into one of four categories: 1=1-10, 2=11-20, 3=21-30, 4=31 or more. Anyone reporting 0 cigarettes per day was coded as missing. This phenotype is a component of commonly used measures of nicotine dependence such as the Fagerstrom Test for Nicotine Dependence.

2. Pack Years (PckYr). Defined in the same way as cigarettes per day but not necessarily binned, divided by 20 (cigarettes in a pack), and multiplied by number of years smoking. This yielded a measure of total overall exposure to tobacco and is relevant to disease outcomes for which smoking is a risk factor, such as cancer and chronic obstructive pulmonary disease risk.

3. Age of Initiation of Smoking (AgeSmk). A measure of early cigarette use. Defined as the age at which a participant first started smoking regularly.

4. Smoking Initiation (SmkInit). A binary variable of whether the individual had ever been a regular smoker (1) or not (0), and often defined as having smoked at least 100 cigarettes during one's lifetime.

5. Drinks per week (DrnkWk). A measure of drinking frequency/quantity. The average number of drinks per week in current or former drinkers.

Genotypes

Fourteen of the seventeen studies were genotyped with the Illumina HumanExome BeadChip, which contains ~250,000 low-frequency nonsynonymous variants, variants from the GWAS catalog, and a small number of variants selected for other purposes. Two studies were genotyped on the Illumina Human Core Exome, which includes an additional ~250,000 tag SNPs. The remaining study, the UK Biobank, was imputed using Haplotype Reference Consortium panel(31, 42), as well as the reference panel by UK 10K and 1000 Genomes Project. An integrated callset was released by the UK Biobank team(42). Our UK Biobank genetic association analyses were conducted based on the integrated callset with additional quality control.

Generation of Summary Association Statistics

Seventeen independent studies (see **Table S1**) with smoking and drinking phenotypes were included in the discovery phase. Individual studies conducted association analysis accounting for age, sex, any study-specific covariates, and ancestry principal components (see **Table S2** for genomic controls), and submitted summary statistics for meta-analysis. For studies with related individuals (**see Table S1**), relatedness was accounted for in linear mixed models

using empirically estimated kinships from common SNPs(43). Residuals were inverse-normalized to help ensure well-behaved test statistics for rare variant tests.

Quality control of per-study summary statistics included evaluation and correction of strand flips and allele flips through systematic comparison of alleles and allele frequencies against the reference datasets ExAC v2.0, 1000 Genomes Phase 3, and dbSNP. Variants with call rates<0.9, or Hardy Weinberg $p<1x10^{-7}$ were also removed. The latter filter was meant to avoid findings that could not be more broadly replicated across the 17 studies.

Meta-analysis

Association testing was done in stages. First, we conducted genome-wide association meta-analysis. Variants with p-values less than the genome-wide significance threshold of 5×10^{-8} were deemed statistically significant. Loci were defined as 1 million basepair windows surrounding a "sentinel" (most significant) variant in the locus. Overlapping or adjacent loci were combined into a single locus. Conditional analysis and fine mapping was then performed within each locus. We attempted to replicate one very rare variant (rs36015615 in *STARD3* associated with CigDay; see results and **Table 1**) that was available in two other exome chip consortia. These were the CHD Exome+ Consortium (N=17,789) and the Consortium for Genetics of Smoking Behaviour (N=28,583). Both consortia defined their phenotypes, including cigarettes per day similarly, as the usual number of cigarettes smoked in a day corrected for sex, age, principal components (and/or genetic relatedness, as appropriate), and inverse-normalized prior to association analysis.

We also conducted gene-level association tests grouping nonsynonymous, stop gain, stop loss and splice variants within each gene, using rareMETALS version 6.0(44). Variant annotation was conducted using SEQMINER with RefSeq 1.9(45). Two complementary gene-level association tests were performed: the sequence kernel association test (SKAT; 46, 47) with a MAF cutoff of 1% and a simple burden test(48) that summed the number of rare alleles within a given gene, again with a maximum MAF=1%. We chose variants with MAF≤1% as we were interested in the contribution of variants with a frequency lower than that which has been reliably

imputed and tested in past GWAS meta-analyses. We considered a gene association to be significant if the p-value surpassed a Bonferroni correction for the number of genes tested for a given phenotype and test, assuming approximately 20,000 genes in the genome (.05/20,000 = 2.5×10^{-6}).

We performed iterative conditional analysis using a partial correlation based score (PCBS) statistic(49), which can perform proper conditional analysis for meta-analysis that combines datasets measured using different arrays. PCBS takes GWAS meta-analysis summary statistics and LD estimated from the Haplotype Reference Consortium panel as input.

As a key step to evaluate the contribution of variants within a genome-wide significant locus(33), we used our PCBS framework to apply two complementary fine mapping techniques to identify putatively causal genetic variants. The first technique was a Bayesian approach described previously(50) that estimates the posterior probability of association based upon the statistical strength of the association for variants in each locus. We also applied a version of fgwas(51) modified to work within PCBS, which assumes that variants in different functional categories have potentially different prior probability of association. For loci with a single association signal based, effect sizes and variance from single-SNP analyses were used. If a locus contained multiple signals, we used effect sizes and variance from conditional analysis adjusting for all other index variants in this region.

Finally, we attempted to replicate previous rare variant associations referenced in the introduction and listed in **Table S4.** We attempted replication in our phenotypes for any single variant when that variant was directly genotyped or imputed. We applied a liberal threshold that corrected only for the number of tests conducted for this replication exercise (.05/46=.001).

Genetic Architecture

We performed heritability and genetic correlation analyses using LD score regression(52). The method calculates LD scores from the Haplotype Reference Consortium and the estimation of heritability with these LD scores then follows established methods(53, 54). Heritability was

estimated for each trait and partitioned by annotation category and frequency bins. First, we annotated variants on the exome chip based upon gene definitions in RefSeq 1.9, using SEQMINER version 6.0(55). A variant is classified as coding if it belongs to either one of the following categories: nonsynonymous, stop gain, stop loss, and splice. Seven functional categories were considered in the model, including intergenic, intron, common coding (MAF>0.01), rare coding (MAF<0.01), synonymous, and 3'/5' untranslated regions. We fitted the baseline model with seven categories, and estimated phenotypic variance explained by each category.

Results

GWAS analyses behaved well, with genomic control values for the GWAS across exome chip and UK Biobank imputed variants between 1.05 and 1.3. The intercept for LD Score regression ranged between .99 and 1.1, indicating absent or minimal effects of population stratification. (Per-study genomic controls can be found in **Table S2**.) A total of 171 loci were identified under the genome-wide significance threshold ($p<5\times10^{-8}$), including 3, 11, 17, 93 and 47 loci for AgeSmk, CigDay, PckYr, SmkInit, and DrnkWk. A list of all sentinel variants within each locus is shown in **Table S5**. QQ plots and Manhattan plots are available in **Figures S1 and S2**. (Additional exploratory GWAS meta-analysis of individuals with significant African ancestry are provided in the Supplementary Note [including up to 8,974 individuals from three studies]; see also **Table S3, Figure S3 and S4.**) The genome-wide significant association results included known loci associated with smoking and alcohol use phenotypes. These included associations between smoking phenotypes and variants within the *CHRNA5-CHRNA3-CHRNB4* nicotinic receptor cluster, nicotine metabolism gene *CYP2A6,* and a locus near dopamine receptor *DRD2*. We also replicated previous associations between nonsynonymous variant rs1229984 in *ADH1B* and DrnkWk. Only one very rare variant was associated with any of our five phenotypes. This was rs36015615 (MAF=.0002), a nonsynonymous variant in *STARD3*, associated with CigDay ($p=3.2\times10^{-8}$). This novel variant did not replicate in either of two replication consortium datasets,

the CHD Exome+ Consortium (N=17,789, Beta=-.01, *p*=.94) or the Consortium for Genetics of Smoking Behaviour (N=28,583, Beta=.056, *p*=.84). Based upon the estimated genetic effects in the discovery sample ($\beta = 1.2$), the power for replication is >99%. However, if we assume the observed effect sizes in the replication datasets are correct, there is 5% power for replication based upon this estimated effect. The pattern of results may be due to winner's curse, or the discovered variant may be a false positive finding. Additional studies are required to narrow the possible interpretations.

The fine mapping analysis of all 171 GWAS loci pinpointed putatively causal variants with high resolution in some cases. The 95% credible interval for 34% of the loci had <10 SNPs and 24 loci had double basepair resolution, including several instances where the sole putative causal variant was nonsynonymous and of lower frequency, although in only one case with MAF<1%. The resolution increased somewhat when functional information was used to inform the prior, with double base-pair resolution at 32 loci, and 44% of loci having <10 SNPs in the 95% credible interval. **Table 1** includes all nonsynonymous or loss of function variants within the genome-wide significant loci that had a posterior probability of association greater than .80 from at least one of the fine mapping methods. Additional results from the fine mapping analysis are available in **Tables S6 and S7**. Several known functional variants were identified through this method, including: rs16969968(56), a nonsynonymous variant in nicotinic receptor gene *CHRNA5* associated with CigDay (PPA=.92 and .84 from the fine mapping analysis with, and without, functional priors, respectively); rs1229984(57), a nonsynonymous variant in alcohol metabolism gene *ADH1B* associated with DrnkWk (PPA=1.0 and 1.0); and, although with somewhat weaker evidence, rs6265(58), a nonsynonymous variant in brain derived neurotrophic factor *BDNF* associated with SmkInit (MAF=.19; PPA=.83 and .32).

Novel variants in novel genes were also prioritized at high resolution. To take the most statistically compelling examples in **Table 1**, we found rs28929474, a low frequency nonsynonymous variant in *SERPINA1*, associated with DrnkWk (MAF=.02; PPA=1.0 and .95).

When homozygous, the alternate T (allele frequency = .02; frequency of TT genotype under Hardy Weinberg = 4 in 10,000) allele is a leading cause of alpha-1 antitrypsin deficiency. Here, we find the same risk allele, the T allele, is associated with an approximately .05 standard deviation decrease in drinks per week. We also discovered rs35891966, a variant in *NAV2*, associated with SmkInit (MAF=.07; PPA = 1.0 and .98) at single base-pair resolution. *NAV2* is involved in neuronal development and previously shown to be differentially expressed between smokers and non-smokers, but not previously implicated in GWAS(59).

Results of gene-based tests are provided in **Table 2**. A novel gene, rho guanine nucleotide exchange factor 37 (*ARHGEF37*), was associated with Age of Initiation of Smoking ($p=1.9\times10^{-6}$). *ARHGEF37* has not been widely studied and its function in not well known. Another novel gene without an immediate biological interpretation, was HEAT Repeat Containing 5A (*HEATR5A*), associated with Smoking Initiation ($p=1.4\times10^{-8}$). We also discovered a significant gene-based association between known alcohol metabolism gene *ADH1C* and Drinks per Week ($p=1.4\times10^{-27}$ and $p=1.9\times10^{-40}$ from the burden and SKAT tests, respectively). Finally, even with relaxed p-value thresholds, we failed to replicate genes identified in previous rare variant association studies referenced in the introduction (**Table S4**), with the exception of *ADH1C* and *CHRNA5*, two loci long known to be associated with alcohol use and smoking, respectively.

The estimated total SNP heritability for AgeSmk, CigDay, PckYr, SmkInit, and DrnkWk was 6%, 9%, 10%, 14% and 16%. Significant phenotypic variance was explained by rare nonsynonymous variants for all traits, ranging from 1.0%-2.2% (**Table 3**). As a fraction of the SNP heritability, rare nonsynonymous variants accounted for 11%-18%. Results for all seven functional categories are listed in **Table S8**; appreciable heritability was accounted for by common and rare coding variants, and intergenic variants. Variants in the untranslated regions and intronic regions contributed less. Almost all pairs of phenotypes were genetically correlated (**Table 4, Panel A**), and the direction of the genetic correlations were in the expected direction. For instance, CigDay was positively correlated with DrnkWk (0.2 ± 0.09) , consistent with the observation that increased

alcohol consumption is correlated with increased tobacco consumption. Age of initiation has a negative correlation with all other traits, which is consistent with the observation that an earlier age of smoking initiation is correlated with increased tobacco and alcohol consumption in adulthood. The patterns and magnitudes of correlation are highly similar when considering only rare nonsynonymous variants (**Table 4, Panel B**).

Discussion

With a maximum sample size ranging from 152,348 to 433,216, the present study is the largest study to date of low-frequency nonsynonymous and loss of function variants in smoking and alcohol use. Our meta-analytic study design combined studies genotyped on the exome array with imputed genotypes in the UK Biobank, allowed us to comprehensively evaluate the contribution of rare and low frequency variants to the etiology of tobacco and alcohol use. All told, we identified 171 genome-wide significant loci for the five phenotypes.

We showed that the rare variants (MAF≤1%) together explain 1.0% - 2.2% of the phenotypic variance for the five traits, amounting to 11-18% of the total SNP heritability. A number of putatively causal low frequency nonsynonymous variants in novel genes were identified through two complementary fine mapping techniques. These include a variant known to affect alpha-1 antitrypsin deficiency in *SERPINA1*. The effect of the risk allele resulted in a decrease in drinks per week. One interpretation is that this variant leads to impaired liver function through alpha-1 antitrypsin deficiency which, in turn, reduces alcohol consumption. Interestingly, neither this particular variant or the locus surrounding it was associated with smoking phenotypes, even though alpha-1 antitrypsin deficiency also affects lung function over time. Other mechanisms by which *SERPINA1* exerts its effect on alcohol consumption are certainly possible. Another novel nonsynonymous variant was in neuron navigator 2 (*NAV2*), associated with smoking initiation. *NAV2* has not previously been associated with substance use or addiction. Given its suspected involvement in neuronal growth and migration, a putatively causal nonsynonymous variant is a

strong candidate for functional follow up experiments. Other genes implicated in the fine mapping analysis have less direct interpretations (e.g., *HEATR5A*) and such results will benefit from replication and/or follow-up experiments. In general, fine mapping studies narrowed the credible set of likely causal variants to single or double base pair resolution for 24 loci (**Table S6**). Some loci were not amenable to fine mapping, with credible intervals containing thousands of SNPs in some cases. Given the cost in money and time of conducting functional experiments at the cellular or organismal level, fine mapping likely causal variants can be extremely useful in predicting functional consequences and prioritizing variants for further work.

Gene based tests identified a small number of associated genes, including an expected association with *ADH1C* and drinks per week. The other two associated genes, *ARHGEF37* and *HEATR5A*, do not lend themselves to ready biological interpretations.

We showed that rare coding variants available on the exome chip or imputable by the Haplotype Reference Consortium, with frequency <1%, explain significant proportions of phenotypic variance, and a substantial proportion of the total SNP heritability. The exome chip was designed to genotype coding variants uncovered in ~12,000 sequenced exomes. By design, it comprehensively ascertained high confidence rare nonsynonymous, splice, and stop variants within those sequences and only sparsely genotypes other classes of variation, including common variants. The Haplotype Reference Consortium panel imputed data also have limited accuracy when the underlying genetic variants are rare. Therefore, our current investigation did not fully explore the genetic architecture of very rare variants (i.e. with MAF<0.1%). With the development of larger imputation reference panels, and the availability of large scale deep whole genome sequences (such as the Trans-Omics for Precision Medicine Study [TOPMed]), we expect to be able to conduct an even more comprehensive analysis of the genetic architecture for variants with ever lower frequencies. Ultimately, the discovery of low frequency with small effects will require even larger sample sizes. For example, for rare variant with MAF of .1% and effects of .2, .15, and 0.1 standard deviations on the phenotype, to identify associations at $\alpha = 5 \times 10^{-8}$ with 80%

of power, sample sizes of 500,000 890,000 and 1,990,000 are required. While such numbers seemed astronomical just a few years ago, they will indeed be attainable in the next few years with the availability of large biobank datasets and ever improving imputation. Another limitation of the present study is the limited samples sizes from non-European ancestries, where only exploratory analyses were possible. Substantial improvements can be made to the resolution of fine mapping analysis by leveraging disparate LD information across samples with diverse ancestry(33). Future research will do well to include individuals of diverse ancestry.

Acknowledgements: Research reported in this article was supported by the National Institute on Drug Abuse and the National Human Genome Research Institute of the National Institutes of Health under award numbers R01DA037904 (SIV), R21DA040177 (DJL), R01HG008983 (DJL) R01GM126479 (DJL) and 5T3DA017637-13 (DMB), as well as funding sources listed in the Supplementary Note. JMH was supported by a NSF Graduate Research Fellowship. This material is based upon work supported by the National Science Foundation Graduate Research Fellowship Program. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Disclosures: There are no conflicts to disclose

References

1. Ezzati M, Lopez AD, Rodgers A, Vander Hoorn S, Murray CJL, Coll CRA (2002): Selected major risk factors and global and regional burden of disease. *Lancet*. 360:1347-1360.
2. Polderman TJ, Benyamin B, de Leeuw CA, Sullivan PF, van Bochoven A, Visscher PM, et al. (2015): Meta-analysis of the heritability of human traits based on fifty years of twin studies. *Nat Genet*.
3. Eng MY, Luczak SE, Wall TL (2007): ALDH2, ADH1B, and ADH1C genotypes in Asians: A literature review. *Alcohol Res Health*. 30:22-27.
4. Furberg H, Kim Y, Dackor J, Boerwinkle E, Franceschini N, Ardissino D, et al. (2010): Genome-wide meta-analyses identify multiple loci associated with smoking behavior. *Nature Genet*. 42:441-U134.
5. Saccone NL, Culverhouse RC, Schwantes-An TH, Cannon DS, Chen X, Cichon S, et al. (2010): Multiple independent loci at chromosome 15q25.1 affect smoking quantity: a meta-analysis and comparison with lung cancer and COPD. *PLoS Genet*. 6.
6. Bierut LJ, Stitzel JA (2014): Genetic Contributions of the alpha 5 Nicotinic Receptor Subunit to Smoking Behavior. *Recept Ser*. 26:327-339.
7. Luczak SE, Glatt SJ, Wall TL (2006): Meta-analyses of ALDH2 and ADH1B with alcohol dependence in Asians. *Psychol Bull*. 132:607-621.
8. Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, et al. (2016): Analysis of protein-coding genetic variation in 60,706 humans. *Nature*. 536:285-+.
9. Minikel E, Lek M, Samocha KE, Karczewski KJ, Marshall JL, Armean I, et al. (2016): An early glimpse of saturation mutagenesis in humans: Insights from protein-coding genetic variation in 60,706 people. *Prion*. 10:S107-S107.
10. Sveinbjornsson G, Albrechtsen A, Zink F, Gudjonsson SA, Oddson A, Masson G, et al. (2016): Weighting sequence variants based on their annotation increases power of whole-genome association studies. *Nat Genet*. 48:314-317.
11. Marouli E, Graff M, Medina-Gomez C, Lo KS, Wood AR, Kjaer TR, et al. (2017): Rare and low-frequency coding variants alter human adult height. *Nature*. 542:186-190.
12. Yang J, Wang S, Yang Z, Hodgkinson CA, Iarikova P, Ma JZ, et al. (2014): The contribution of rare and common variants in 30 genes to risk nicotine dependence. *Mol Psychiatry*.
13. McClure-Begley TD, Papke RL, Stone KL, Stokes C, Levy AD, Gelernter J, et al. (2014): Rare human nicotinic acetylcholine receptor alpha4 subunit (CHRNA4) variants affect expression and function of high-affinity nicotinic acetylcholine receptors. *The Journal of pharmacology and experimental therapeutics*. 348:410-420.
14. Piliguian M, Zhu AZ, Zhou Q, Benowitz NL, Ahluwalia JS, Sanderson Cox L, et al. (2014): Novel CYP2A6 variants identified in African Americans are associated with slow nicotine metabolism in vitro and in vivo. *Pharmacogenet Genomics*. 24:118-128.
15. Haller G, Druley T, Vallania FL, Mitra RD, Li P, Akk G, et al. (2012): Rare missense variants in CHRNB4 are associated with reduced risk of nicotine dependence. *Hum Mol Genet*. 21:647-655.
16. Haller G, Kapoor M, Budde J, Xuei X, Edenberg H, Nurnberger J, et al. (2014): Rare missense variants in CHRNB3 and CHRNA3 are associated with risk of alcohol and cocaine dependence. *Hum Mol Genet*. 23:810-819.
17. Zuo L, Tan Y, Li C-SR, Wang Z, Wang K, Zhang X, et al. (2016): Associations of rare nicotinic cholinergic receptor gene variants to nicotine and alcohol dependence. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*.
18. Xie P, Kranzler HR, Krauthammer M, Cosgrove KP, Oslin D, Anton RF, et al. (2011): Rare Nonsynonymous Variants in Alpha-4 Nicotinic Acetylcholine Receptor Gene Protect Against Nicotine Dependence. *Biological Psychiatry*. 70:528-536.

19. Wessel J, McDonald SM, Hinds Da, Stokowski RP, Javitz HS, Kennemer M, et al. (2010): Resequencing of Nicotinic Acetylcholine Receptor Genes and Association of Common and Rare Variants with the Fagerström Test for Nicotine Dependence. *Neuropsychopharmacology*. 35:2392-2402.
20. Thorgeirsson TE, Steinberg S, Reginsson GW, Bjornsdottir G, Rafnar T, Jonsdottir I, et al. (2016): A rare missense mutation in CHRNA4 associates with smoking behavior and its consequences. *Molecular Psychiatry*. 21:594-600.
21. Olfson E, Saccone NL, Johnson EO, Chen L-S, Culverhouse R, Doheny K, et al. (2016): Rare, low frequency and common coding variants in CHRNA5 and their contribution to nicotine dependence in European and African Americans. *Molecular Psychiatry*. 21:601-607.
22. Peng Q, Gizer IR, Libiger O, Bizon C, Wilhelmsen KC, Schork NJ, et al. (2014): Association and ancestry analysis of sequence variants in ADH and ALDH using alcohol-related phenotypes in a Native American community sample. *Am J Med Genet B Neuropsychiatr Genet*. 165B:673-683.
23. Way M, McQuillin A, Saini J, Ruparelia K, Lydall GJ, Guerrini I, et al. (2015): Genetic variants in or near ADH1B and ADH1C affect susceptibility to alcohol dependence in a British and Irish population. *Addiction Biology*. 20:594-604.
24. Zuo L, Wang K-S, Zhang X-Y, Li C-sR, Zhang F, Wang X, et al. (2013): Rare SERINC2 variants are specific for alcohol dependence in individuals of European descent. *Pharmacogenetics and Genomics*. 23:395-402.
25. Zuo L, Zhang X, Deng H-w, Luo X (2013): Association of rare PTP4A1-PHF3-EYS variants with alcohol dependence. *Journal of Human Genetics*. 58:178-179.
26. Haller G, Li P, Esch C, Hsu S, Goate AM, Steinbach JH (2014): Functional Characterization Improves Associations between Rare Non-Synonymous Variants in CHRNB4 and Smoking Behavior. *PLoS ONE*. 9:e96753.
27. Duncan LE, Keller MC (2011): A Critical Review of the First 10 Years of Candidate Gene-by-Environment Interaction Research in Psychiatry. *Am J Psychiat*. 168:1041-1049.
28. Vrieze SI, Feng S, Miller MB, Hicks BM, Pankratz N, Abecasis GR, et al. (2014): Rare nonsynonymous exonic variants in addiction and behavioral disinhibition. *Biol Psychiatry*. 75:783-789.
29. Vrieze SI, Malone SM, Vaidyanathan U, Kwong A, Kang HM, Zhan X, et al. (2014): In search of rare variants: preliminary results from whole genome sequencing of 1,325 individuals with psychophysiological endophenotypes. *Psychophysiology*. 51:1309-1320.
30. Vrieze SI, Malone SM, Pankratz N, Vaidyanathan U, Miller MB, Kang HM, et al. (2014): Genetic associations of nonsynonymous exonic variants with psychophysiological endophenotypes. *Psychophysiology*. 51:1300-1308.
31. McCarthy S, Das S, Kretzschmar W, Delaneau O, Wood AR, Teumer A, et al. (2016): A reference panel of 64,976 haplotypes for genotype imputation. *Nature genetics*. 48:1279-1283.
32. Evans LM, Tahmasbi R, Vrieze SI, Abecasis GR, Das S, Gazal S, et al. (2018): Comparison of methods that use whole genome data to estimate the heritability and genetic architecture of complex traits. *Nature Genet*. 50:737-+.
33. Schaid DJ, Chen W, Larson NB (2018): From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nat Rev Genet*. 19:491-504.
34. Hicks BM, Schalet BD, Malone SM, Iacono WG, McGue M (2011): Psychometric and genetic architecture of substance use disorder and behavioral disinhibition measures for gene association studies. *Behav Genet*. 41:459-475.
35. Vrieze SI, McGue M, Miller MB, Hicks BM, Iacono WG (2013): Three mutually informative ways to understand the genetic relationships among behavioral disinhibition, alcohol use, drug use, nicotine use/dependence, and their co-occurrence: twin biometry, GCTA, and genome-wide scoring. *Behav Genet*. 43:97-107.

36. Vink JM, Willemsen G, Boomsma DI (2005): Heritability of smoking initiation and nicotine dependence. *Behav Genet*. 35:397-406.
37. Maes HH, Sullivan PF, Bulik CM, Neale MC, Prescott CA, Eaves LJ, et al. (2004): A twin study of genetic and environmental influences on tobacco initiation, regular tobacco use and nicotine dependence. *Psychological Medicine*. 34:1251-1261.
38. Swan GE, Carmelli D, Rosenman RH, Fabsitz RR, Christian JC (1990): Smoking and alcohol consumption in adult male twins: genetic heritability and shared environmental influences. *J Subst Abuse*. 2:39-50.
39. Schumann G, Liu CY, O'Reilly P, Gao H, Song P, Xu B, et al. (2016): KLB is associated with alcohol drinking, and its gene product beta-Klotho is necessary for FGF21 regulation of alcohol preference. *P Natl Acad Sci USA*. 113:14372-14377.
40. Jorgenson E, Thai KK, Hoffmann TJ, Sakoda LC, Kvale MN, Banda Y, et al. (2017): Genetic contributors to variation in alcohol consumption vary by race/ethnicity in a large multi-ethnic genome-wide association study. *Mol Psychiatry*.
41. Thorgeirsson TE, Gudbjartsson DF, Surakka I, Vink JM, Amin N, Geller F, et al. (2010): Sequence variants at CHRNB3-CHRNA6 and CYP2A6 affect smoking behavior. *Nature Genet*. 42:448-U135.
42. Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, et al. (2017): Genome-wide genetic data on ~500,000 UK Biobank participants. *bioRxiv*.
43. Kang HM, Sul JH, Service SK, Zaitlen NA, Kong SY, Freimer NB, et al. (2010): Variance component model to account for sample structure in genome-wide association studies. *Nat Genet*. 42:348-354.
44. Liu DJ, Peloso GM, Zhan X, Holmen OL, Zawistowski M, Feng S, et al. (2014): Meta-analysis of gene-level tests for rare variant association. *Nat Genet*. 46:200-204.
45. Pruitt KD, Brown GR, Hiatt SM, Thibaud-Nissen F, Astashyn A, Ermolaeva O, et al. (2014): RefSeq: an update on mammalian reference sequences. *Nucleic Acids Res*. 42:D756-763.
46. Wu MC, Lee S, Cai TX, Li Y, Boehnke M, Lin XH (2011): Rare-Variant Association Testing for Sequencing Data with the Sequence Kernel Association Test. *Am J Hum Genet*. 89:82-93.
47. Lee S, Wu MC, Lin X (2012): Optimal tests for rare variant effects in sequencing association studies. *Biostatistics*. 13:762-775.
48. Li B, Leal SM (2008): Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am J Hum Genet*. 83:311-321.
49. Jiang Y, Chen S, McGuire D, Chen F, Liu M, Iacono WG, et al. (2018): Proper conditional analysis in the presence of missing data: Application to large scale meta-analysis of tobacco use phenotypes. *PLoS genetics*. 14:e1007452.
50. Mahajan A, Taliun D, Thurner M, Robertson NR, Torres JM, Rayner NW, et al. (2018): Fine-mapping of an expanded set of type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps. *bioRxiv*.
51. Pickrell JK (2014): Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. *Am J Hum Genet*. 94:559-573.
52. Bulik-Sullivan BK, Loh PR, Finucane HK, Ripke S, Yang J, Schizophrenia Working Group of the Psychiatric Genomics C, et al. (2015): LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nature genetics*. 47:291-295.
53. Bulik-Sullivan BK, Loh PR, Finucane HK, Ripke S, Yang J, Patterson N, et al. (2015): LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nature Genet*. 47:291-+.
54. Bulik-Sullivan B, Finucane HK, Anttila V, Gusev A, Day FR, Loh PR, et al. (2015): An atlas of genetic correlations across human diseases and traits. *Nature Genet*. 47:1236-+.

55. Zhan X, Liu DJ (2015): SEQMINER: An R-Package to Facilitate the Functional Interpretation of Sequence-Based Associations. *Genet Epidemiol*.
56. Lassi G, Taylor AE, Timpson NJ, Kenny PJ, Mather RJ, Eisen T, et al. (2016): The CHRNA5-A3-B4 Gene Cluster and Smoking: From Discovery to Therapeutics. *Trends Neurosci*. 39:851-861.
57. Edenberg HJ (2007): The genetics of alcohol metabolism: role of alcohol dehydrogenase and aldehyde dehydrogenase variants. *Alcohol Res Health*. 30:5-13.
58. Ghitza UE, Zhai H, Wu P, Airavaara M, Shaham Y, Lu L (2010): Role of BDNF and GDNF in drug reward and relapse: a review. *Neurosci Biobehav Rev*. 35:157-171.
59. Paul S, Amundson SA (2014): Differential Effect of Active Smoking on Gene Expression in Male and Female Smokers. *J Carcinog Mutagen*. 5.

Table 1. All nonsynonymous/loss of function variants with posterior probability of association > .80 from one of the two fine mapping methods.

SNP	REF/ALT	N	ALT AF	GWAS p-value	Beta	SE	Direction	Annotation	Posterior Probability of Association		Number SNPs (Low Frequency Coding SNPs) in 95% Credible Interval	
									W/out Functional Prior	W/ Functional Prior (fgwas)	W/out Functional Prior	W/ Functional Prior (fgwas)
Cigarettes per Day (CigDay)												
rs36015615[a]	G/A	69,951	.0002	3.2×10^{-8}	1.2	.210	==+=++=+=X=X+X+++	Nonsynonymous [*STARD3*]	.82	.62	8,997 (6211)	11302 (6232)
rs16969968	G/A	153,918	.34	2.5×10^{-139}	.096	.0038	+-++-+++++-++++++	Nonsynonymous [*CHRNA5*]	.84	.92	2(0)	2 (0)
Drinks per Week (DrnkWk)												
rs1260326	T/C	357,854	.61	4.6×10^{-40}	0.032	.0024	+++++++++--++++++	Nonsynonymous [*GCKR*]	1.0	1.0	1 (0)	1 (0)
rs1229984	T/C	334,588	.98	2.3×10^{-173}	0.25	.0088	=+-XXXX+XXXX=++++	Nonsynonymous [*ADH1B*]	1.0	1.0	1 (1)	1 (1)
rs28929474	C/T	357,854	.02	2.2×10^{-11}	-0.057	.0085	--++-----+++----+	Nonsynonymous [*SERPINA1*]	.95	1.0	1 (1)	1 (1)
rs1800566	G/A	357,854	.18	2.00×10^{-8}	0.017	.0031	++++++++++--++++-	Nonsynonymous [*NQO1*]	.32	.97	103 (0)	1 (0)
Smoking Initiation (SmkInit)												
rs2232423	A/G	433,216	.11	1.40×10^{-8}	-0.019	.0034	-+-+---+------	Nonsynonymous [*ZSCAN12*]	.84	.64	502 (0)	2 (0)
rs35891966	G/A	433,216	.07	1.30×10^{-8}	-0.024	.0042	---+-+---++---	Nonsynonymous [*NAV2*]	.98	1.0	1 (0)	1 (0)
rs147052174	G/T	433,216	.02	1.2×10^{-7}	.043	.0080	+++++++++-++++	Nonsynonymous [*FAM163A*]	.81	1.0	2432(66)	1 (0)
rs6265	C/T	433,216	.19	1.9×10^{-10}	-.017	.0030	++-++-++--+---	Nonsynonymous [*BDNF*]	.32	.83	25(0)	2 (0)
rs61754158	C/T	433,216	.01	1.4×10^{-6}	-.055	.0114	--+-+---=+=-+-	Nonsynonymous [*HEATR5A*]	.39	.87	9742(195)	9742 (195)
rs34967813	A/G	433,216	.31	8.1×10^{-7}	-.011	.0023	----+--+--+---	Nonsynonymous [*RYR2*]	.14	.98	7413(56)	1 (0)

[a]rs36015615 did not replicate in two additional datasets. See results section.

Note: REF=reference allele on GRCh37, ALT=alternate allele, N=sample size across all studies that genotyped the variant, ALT AF=allele frequency of the alternate allele estimated in the meta-analysis. A variant is considered “rare” if MAF < .01, and low frequency if .01 ≤ MAF < .05. In the DIRECTION column each symbol represents the contribution of one of the studies. A “+” indicates the ALT allele had a positive effect in that study; “-“ indicates a negative effect. A “=” indicates the variant was monomorphic and “X” indicates it was absent in that study. The order of studies for CigDay and DrnkWk was ARIC, UKB, COGA, FINNTWIN, FUSION, GECCO, HRS, ID1000, MEC, METSIM, MHI, MCTFR, NAGOZALC, NESCOG, SardiNIA, TwinsUK, and WHI. For SmkInit the order is the same except COGA and MCTFR were not available. See the supplemental materials for study acronym explanations.

Table 2. Significant gene based test results, assuming a Bonferroni threshold of .05/20,000=2.5×10^{-6}.

Phenotype	Gene	N	Number Variants	Beta	SE	p-value	Method
Age of Initiation of Smoking	*ARHGEF37*	147,010	17	.08	.017	1.9×10^{-6}	Burden
Smoking Initiation	*HEATR5A*	427,262	41	-.02	.009	1.4×10^{-8}	SKAT
Drinks per Week	*ADH1C*	353,265	4	-.15	.014	1.8e-27	Burden
Drinks per Week	*ADH1C*	353,265	4	-.15	.014	1.9e-40	SKAT

Note: no significant genes were identified for the other two phenotypes.

Table 3: Estimation of Heritability Explained by Variants on Exome Array. We estimate the heritability based upon a baseline model with seven different functional categories. The reported heritability $\hat{h}^2$ is based upon the cumulative value from the functional categories with significant heritabilities. We also report the standard deviation ($se(\hat{h}^2)$) and p-values, estimated using jackknife.

Annotation	**Phenotype**	**Heritability Estimates**		
		$\widehat{\boldsymbol{h}}^{\boldsymbol{2}}$	$\boldsymbol{se}(\widehat{\boldsymbol{h}}^{\boldsymbol{2}})$	**P-Value**
All Variants	Age of Initiation of smoking	.06	.0049	7.7×10^{-35}
	Cigarettes per Day	.09	.0019	$< 2.2\times10^{-303}$
	Pack Years	.10	.0022	$< 2.2\times10^{-303}$
	Smoking Initiation	.14	.0007	$< 2.2\times10^{-303}$
	Drinks per Week	.16	.0089	7.3×10^{-73}
Rare Coding Variants (MAF<.01)	Age of Initiation of smoking	.011	.0015	2.8×10^{-2}
	Cigarettes per Day	.010	.0006	1.7×10^{-2}
	Pack Years	.018	.0007	8.5×10^{-6}
	Smoking Initiation	.022	.0002	3.9×10^{-16}
	Drinks per Week	.020	.0013	1.8×10^{-7}

Table 4: Estimation of Genetic Correlation Between Smoking and Drinking Traits. We estimate genetic correlations between five smoking and drinking traits. Genetic correlation estimates ($\hat{r}_g$), their standard deviation ($se(\hat{r}_g)$) and p-values are reported.

Trait 1	Trait 2	Genetic Correlation		
		$\hat{r}_g$	$se(\hat{r}_g)$	P-value
A. Aggregated Genetic Correlation Induced by All Variants on the Exome Array				
Drinks per Week	Smoking Initiation	.43	.06	1.7×10^{-11}
Drinks per Week	Age of Initiation of Smoking	.01	.13	9.3×10^{-1}
Drinks per Week	Pack Years	.22	.10	2.6×10^{-2}
Drinks per Week	Cigarettes per Day	.20	.09	3.1×10^{-2}
Smoking Initiation	Age of Initiation of Smoking	-.64	.11	1.1×10^{-8}
Smoking Initiation	Pack Years	.45	.08	4.9×10^{-8}
Smoking Initiation	Cigarettes per Day	.10	.07	1.5×10^{-1}
Age of Initiation of Smoking	Pack Years	-.63	.17	2.1×10^{-4}
Age of Initiation of Smoking	Cigarettes per Day	-.26	.16	9.9×10^{-2}
Pack Years	Cigarettes per Day	.77	.13	2.2×10^{-9}
B. Genetic Correlation Induced by Rare (MAF < 1%) Nonsynonymous Variants				
Drinks per Week	Smoking Initiation	.49	.08	1.2×10^{-10}
Drinks per Week	Age of Initiation of Smoking	-.04	.30	8.9×10^{-1}
Drinks per Week	Pack Years	.08	.02	2.7×10^{-4}
Drinks per Week	Cigarettes per Day	.09	.02	5.2×10^{-5}
Smoking Initiation	Age of Initiation of Smoking	-1.10	.21	1.3×10^{-7}
Smoking Initiation	Pack Years	.63	.08	1.5×10^{-14}
Smoking Initiation	Cigarettes per Day	.23	.08	3.3×10^{-3}
Age of Initiation of Smoking	Pack Years	-1.10	.33	1.5×10^{-3}
Age of Initiation of Smoking	Cigarettes per Day	-.69	.32	3.2×10^{-2}
Pack Years	Cigarettes per Day	.87	.14	1.4×10^{-9}